Philosophical Papers
Vol. 37, No. 2 (July 2008): 289-317

On Planning: Toward a Natural History of Goal Attainment

Mariam Thalos

Abstract: The goal of the essay is to articulate some beginnings for an empirical approach to the study of agency, in the firm conviction that agency is subject to scientific scrutiny, and is not to be abandoned to high-brow aprioristic philosophy. Drawing on insights from decision analysis, game theory, general dynamics, physics and engineering, this essay will examine the diversity of planning phenomena, and in that way take some steps towards assembling rudiments for the budding science, in the process innovating (parts of) a technical vocabulary. The key is focus upon the organization of effort in time. This paper categorizes forms of organization of effort in time, and yields an analysis of *both* individual agency *and* coalitions of agents *as* forms of effort organized in time. Finally, it articulates precise questions pertaining to the natural (evolutionary) history of forms of agency (once upon a time referred to as 'Will') that we now find on the ground.

Introduction

Planning has recently come in for considerable philosophical approbation, indeed as something an entity must perform in order to qualify as fully rational. (This is collateral to something that has been accepted for some time: that having projects, and especially long-term projects, renders an entity worthy of more—or at any rate of different types of—moral consideration, particularly when it comes to matters of public policy.) Facets of planning have come to be viewed as marks of rationality itself, if not of morality too, bearing (or so they have been saying recently) on whether it is ever rational (and furthermore advisable) to make promises and issue threats. I shall argue here that the case for planning has been inflated, fundamentally as a result of too grandiose a conception of its place (or more precisely, lack of it) in nature: it is instead misconceptions of planning as a natural phenomenon that have wrought the highfalutin language that carries with it the favored morals about rationality in the first instance, and

accordingly the preferred doctrines about due moral consideration. For planning, as a matter of cognitive activity whose function is to attain a goal, stands at one end of a spectacularly diverse but otherwise quite orderly continuum of natural possibilities for self-organization, many of them quite humble. Collecting all these under one schema or taxon for systematic study will be our task here, undertaken in service of eventually illuminating the natural history of Will, without wrenching it stillborn from its natural context and casting upon it the fluorescent light of high-brow analysis.

The study of planning is still in its infancy, with some roots in (on one side) decision analysis, game theory, general dynamics, and computer science, and other roots (on another side) in behavioral ecology and other social sciences. This essay will take some steps towards assembling rudiments for this budding science, innovating (parts of) a technical vocabulary, but drawing on ideas that can be found already in physics, engineering, economics and decision analysis. The key is to focus upon the organization of effort in time. This paper categorizes forms of organizing effort in time, and yields an analysis of *both* individual agency *and* coalitions of agents *as* forms of effort organized in time. Finally, it articulates precise questions pertaining to the natural (evolutionary) history of forms of agency (once upon a time referred to as 'Will') that we now find on the ground.

The goal of the essay is to articulate some beginnings for an empirical approach to the study of agency, firmly committed to the idea that agency is subject to scientific scrutiny, and not to be abandoned to high-brow aprioristic philosophy. Aprioristic analyses of agency systematically fail to take natural facts about organismic development, risk and distribution of control over numerous sites of behavior seriously. And while the instincts of some scientifically-minded theorists (predominantly economists) would be to try to make do without a metaphysic of agency, this project will aim at a much more sophisticated—and, more

importantly, empirically-minded—metaphysic, alert to a full range of natural possibilities within a realistic evolutionary setting.[1]

Navigating a dicey world

When an assembly—whether of many, or simply an assembly of one—convenes to devise a plan of action in relation to some novel situation, or vis-à-vis some ongoing aspect of life rife with risk, that assembly is recognizing the call for action. This ability to recognize the call for action is an aspect of practical wisdom. To be practically wise is to exercise sensitivity to guidance—an appreciation of features of the world that call for specific behaviors—in such a way as serves a purpose, goal or end. The end may be one that the wise one aspires to inherently (through organic necessity, perhaps), or an end that the wise one comes to aspire to *through* possessing a susceptibility to guidance. (This latter is, in common parlance, the deliberate activity of setting goals). But what exactly does susceptibility to guidance amount to? I will refer to this as the capacity for *navigation*.

What is navigation? Hunting behaviors of predators from every phylum of the animal kingdom certainly qualify as inspiring pieces of navigation—more so when these predators work in packs. But then so do the many foraging and predator-evading behaviors of herbivores. Some maneuvering in relation to predators, on the part of prey, draw attention in a big way, on account of being elaborately organized, and are successful at least partly because they do. This is the provenance of animal herding. Consider for example the flocking of birds. Throughout his life, field naturalist Edward Selous struggled to explain the astonishing synchrony and coherence of motion in a starling flock, a swarming mass of insects, or a sweeping, twisting school of tiny silvered fish. He concluded in 1905 that '[t]hey must think collectively, all at the same time, or at least in streaks or patches—a square yard or so of an

1 I am grateful to Don Ross, both for the suggestion of the roadmap, and for (roughly) this way of characterizing the way on it.

idea, a flash out of so many brains' (as quoted by Couzin, 2007). It would be easy to conclude that, for lack of more scientifically respectable notions, Selous yielded to the seductive temptations of the disreputable idea of a collective mind, a Victorian notion imbued with immoderate romantic fascination. But this judgment would do a disservice to the insights towards which he was groping. As we will eventually put it, birds in formation (N of them, say) have reduced their *degrees of freedom* from some multiple of N to something considerably smaller. And this is the navigational marvel that Selous sought to grasp. How is this marvel of navigation accomplished?

A starting thought is that only organisms with mobility—with the capacity for displacing themselves in space—can benefit from or display an ability to navigate. But this is not obviously the case. A Venus Flytrap plant might not be navigating in this most robust sense of the term. Still, it is taking some initiative in service of its survival: it 'prepares' a trap well in advance and waits for it to be triggered. Then it takes appropriate 'action' to process its catch when that action is called for. (Similar remarks will apply to a human being who suffers from pervasive paralysis.)

Navigation is thus not easily demarcated. Still, one wants to say that an organism cannot navigate if it can take no initiative of *any* kind. But what, in turn, is initiative? I shall for purposes of this study say that initiative is a matter of having one's behavior organized along a timeline so that it serves a goal. The key here is 'organized'; this qualifier ensures that meeting the goal in question is not merely an accidental happenstance. But the notion of organization does not require that the behavior in question be premeditated in keeping with some soap-operatic paradigm—it does not require a capacity for that mental activity of reasoning in means-ends style that can under the right conditions earn an entity the label of 'devious.' Indeed it may require no mental activity of any kind. If successful, therefore, this account will amount to (at least the rudiments of) a theory of navigation, conceived as *behavior organized in service of a goal*. The principles of analysis will apply to all

organisms with interests to promote, from the most non-deliberating to the most deliberate about their goals. In other words, it shall constitute the fundaments of a very general theory of goal attainment, and therefore of agency, that applies to the deliberate and the indeliberate alike.

Navigating in time

There are two important and distinct things that go into the exercise of initiative. The first is the *control scheme* under which effort is exerted, and which gives the initiative an authorial structure—a structure that allows us to assign, in as fine and feature-sensitive a fashion as we insist upon, responsibility for actions, contrivances and collusions. We shall not devote much space here to illuminating the conceptual underpinnings of this feature, though we shall have a few words to say about how authorial structure interacts with the second feature of the exercise of initiative.[2] This second feature is the *actual shape that the effort itself takes in time*. What we shall undertake here is illumination of this idea of 'the shape of effort in time'. This idea is one that is discussed by professionals in the disciplines of management and long-term planning, and involves the framing of, and taking steps toward meeting of goals. Discussions in management texts are lacking in rigor. We shall here attempt to supply some of that rigor.

Meeting goals through investment of effort is a matter of working according to a timeline or schedule, and in such a way as is sensitive to developments that impact task completion. And so we require looking carefully at how goals are met along a timeline or schedule in such a manner. What is it to be so organized *as to serve or enhance the attainment of goals on a schedule*? This is the question we shall endeavor to assemble tools for answering. It will turn out that there are a number (probably quite a small number) of ways of being so organized. These will

2 A fuller account of authorial issues can be found in Thalos (2007), which builds upon Norman and Shallice (1986) and Hardcastle (1995).

correspond to ways of organizing one's resources as an agent—indeed, they correspond to ways of organizing *oneself* as an agent. (Hence the few words that must be said about authorial structure along the way.) They are thus ways of agenthood. These ways will include organization as a coalition of separate individuals, not necessarily even all belonging to the same species.

Consider an early specimen of *homo sapiens*, after a juicy bit of rabbit with a rock, spear or bow and arrow. An even earlier specimen or ancestor of the species might not have gone after fast-moving targets, and was just content to wander in search of roots, berries and easy prey. Upon the specimen we are now considering—the more evolved specimen—the planning that goes into successful projectile hunting surely confers fitness advantages. Hunting requires planning because once a projectile has been launched, there can be no correction for errors—no adjustment to developments in progress. If the arrow fails its mark, the quarry is lost, and with it a hefty investment of time and precious metabolic energy. Someone might nonetheless insist that planning is not involved in the hurling of a projectile; it is all simply a matter of honing largely unconscious skill. Never mind. The only point I am making is that this skill, conscious or otherwise, that goes into projectile hunting—whatsoever we choose to call it—is different from what is involved in an episode of hunting the quarry down on foot or by hand (drawing on raw brawn and speed), different still from what is involved in extracting and preparing nutritious but initially poisonous roots, and different still yet from what is involved in cultivation of land for production of hand-selected crops. For in the latter, contrasting cases, *feedback* is available throughout the process, in such a way as allows adjustments (although in the last case the time frame is so long as to call for a separate analysis of the skills involved). There is time enough to adjust for errors made, or unforeseen turns of events. But in the case of an arrow sent through the air, no adjustment is possible.

Hunting with a projectile is a matter of managing one's efforts in such a way that potential infelicities of every sort are catered for before

they actually arise. It is a matter of packaging one's efforts in advance. This *front-loading* technique or capacity—as I propose we refer to it—is a means of compensating for the unavailability (or, more precisely, the unusability) of feedback. And as such this capacity is thorough-going future oriented. Small wonder it serves as the cinematic paradigm of planning. And it requires a large and special metabolic investment in the capacities—cognitive and otherwise—that render it possible.

Of course, as what we've said already will have already indicated, not all of the ways of being organized conforms to the front-loading pattern. Some, like that involved in agriculture, is required because the time interval between the time of inauguration or investiture, and the time the goal is reached, is so large. What's more, there's always an opportunity to modify the plan: feedback is available, as well as the *opportunity* for large or minute adjustments. A plan, in these instances, will be your present self's way of coordinating with past and future selves. I propose to refer to this pattern as the *coordination* pattern.

This is essentially the difference between the philosopher Michael Bratman's account of planning (1987, 1999, 2000a), and that of David Gauthier (1994, 1997). Bratman's account is front-loading, inspired by the future orientation model exemplified in projectile hunting. Gauthier's coordination account is by contrast focused upon examples where planning is conducted as a means of coordinating with all of one's temporal parts, past and future, in a way that is open to numerous adjustments along the way, as a way of taking advantage of the benefits—as well as a way of compensating for the disadvantages—of having through no fault of one's own to distribute one's agentic efforts across time. Gauthier is thus able to take advantage of (or, more honestly, create) the opportunities for precommitment where feasible, whilst these advantages weigh in much less heavily for Bratman.

Obviously, successful navigation in a given circumstance might require one or the other of these planning strategies, either exclusively or each in correct measure: some cases require more front loading, others more coordination of past and future selves, and yet others only

one but not the other. And so a skilled navigator (that ideal, possibly purely imaginary organism), equipped with both capacities, has to select a strategy that befits the circumstance, whilst also seeking to compensate for the disadvantages inherent in having to make a selection at the point of need in the first place (for example, the costs in time and cognitive resources). Once—if ever—a capacity for strategy selection (whatever term we wish to use for this, additional, executive function) comes on the scene, it will be called upon—and stretched—to perform such work ever more efficiently. And upon cases where an agent simply *cannot* be *both* future-oriented *and* compliantly coordinative with all of one's selves in time—theories that identify practical wisdom with just one or the other of these two functions, will of course end up in disagreement amongst themselves.

The surviving question, once we've acknowledged all this, must be: *Can* the demands between the potentially diverging strategies of pure future-orientation, on the one hand, and that of coordination, on the other, be reconciled—or, better yet, combined into a higher-order strategy? And if they cannot be combined, can there be an even higher-ranking planning function—an executive of some sort—that selects between them on a case-by-case basis, in a cost-effective way? If so, what does that value-adding executive function *itself* look like? If not, how can a balance be struck in advance of seeing the actual cases of choice on the ground?

I do not pretend any answers to these questions of substance here. But before we leave the matter, we must also examine whether there might not be other strategies that recommend themselves to the practically wise. If there is front loading, might there not also be *back-loading*, a strategy in which one commits to an action path whose full outcome is partly determined by at least one further action taken by the agent or turn of events down the road? (Numerous real-life contracts—like, for example, wills—work along these lines.) This back-loading strategy is somewhere intermediate between front-loading and the coordinative strategy, and whether it is available in real-life situations is

really an empirical matter—a matter of whether agents extended in time can avail themselves of 'contracts' of the right sort, in the ecological context, and if not, whether these things can be self-engineered by the agents who would avail themselves of them.

We shall not go in search of further ways of organizing agency, but will leave further examinations to others. Still, it bears mentioning that these ways will vary according to whether multi-tasking is available, whether tasks and their processing are distributed across sites of agency, and (if distributing is chosen) how processing is actually distributed. And so there is an important point to be made in connection with the question of patterns of organizing agency: it is interlinked with the question of how the claims and duties of self, are ranked against the claims of groups to which one belongs, and even against the possibly diverging claims of sub-personal parts or features of a single self (one's knees versus one's heart, say, in the matter of whether and how to engage in physical exercise). The question of how agency is organized therefore interlinks with the question of whether there is a unit of agency, amongst those entities interacting in any given instance, whose ends or welfare must be ranked highest (and further, how that ranking itself is accomplished). And this issue is sure to have bearing on the question: among all the shapes that planning can take, is one to be preferred—practically, morally, or otherwise—to all the others?

The preponderance of philosophical thinking, conducted in the aprioristic style characteristic or our philosophical era, would suggest that the answer to this last question must be an affirmative. But a naturalistic or empiricist stance would advise that we examine whether the choice among options might be entirely a matter of ecology: if circumstances are such that front-loading pays better or plays better, then we should find a preponderance of front-loading among agents in nature. And if neither is universally better, we should expect a population of agents utilizing different forms—and mixes—of the different strategies.

The formal apparatus

The remainder of my remarks require formulating an interconnected set of technical terms. These terms will allow us to articulate an account of goal attainment: *opportunity structure, choice, dilemma, action point/node, action path* and *outcome*. I will say that action is called for at certain *action points or nodes*, that these are organized in *structures of opportunity*, that an agent makes a *choice* in a *dilemma* when the action point in which that agent finds herself admits of more than one *outcome*, and that the actual outcome depends upon which *action path* the agent takes.

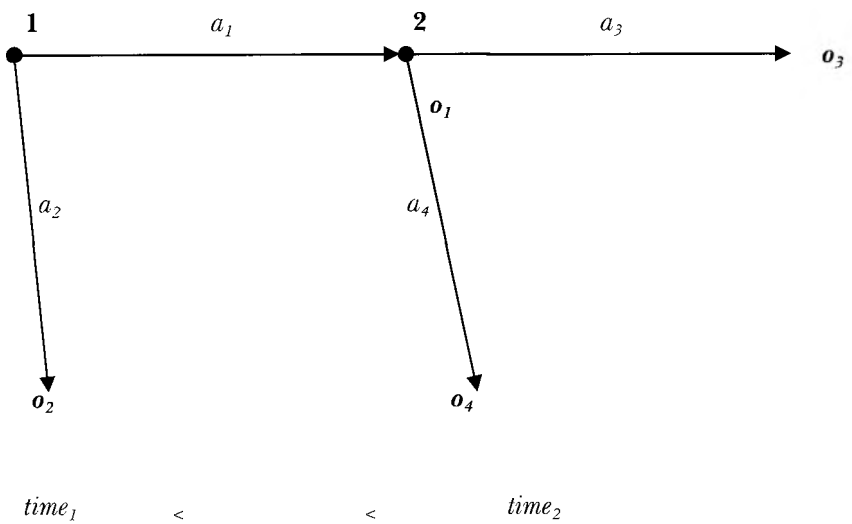

Figure 1: A hypothetical structure ($T_{1\text{-}2}$)

Let us refer to structures of opportunities also as *decision trees*, and name a tree by a sequence of (numbered) action or decision points or nodes, corresponding to a time-indexed sequence of opportunities for choice. The choice situation depicted abstractly in Figure 1 is therefore $T_{1\text{-}2}$. On this way of referring to decision trees, T_2 is another tree, depicting an imbedded opportunity or dilemma, and is therefore related to $T_{1\text{-}2}$ as follows: $T_2 \subset T_{1\text{-}2}$. Let us refer to actions or action

paths chosen (or deserving to be) as *solutions;* the solution sets of $T_{1\text{-}2}$ we can denominate as '$S(T_{1\text{-}2})$,' making no commitment as to whether that set contains a unique element. An element of a solution set will be an action path, consisting of a sequence of actions a_i - a_j - ... - a_n. Let us say that $S(T) \subset S(R)$ —that is to say that the solutions overlap—when two conditions hold: (1) $T \subset R$, and (2) sequence S(T) matches the relevant part of sequence S(R).

Principles and solution concepts

Here are two principles for identifying genuine solutions to navigation problems conceived simply as means-ends problems. Each has been endorsed, in one way or another, in formal decision theoretic literature.

1. *Principle of maximal outcome:* When choosing between two or more paths of action at a node **n**, the only relevant features of the choice are the sequences of outcomes $\boldsymbol{o}_i$ associated with each action path future to **n**.

We might think of this as a form of consequentialism. This principle is another way of putting the point that the solution concept we are after here is a solution concept to a problem conceived purely as a means-ends problem. And it advises that we aim at the best option when taking action.

A second principle goes like this:

2. *Principle of separable choice:* When choosing at the initial node between courses of action that involve taking subordinate actions at different points in time (for example, between a_i - a_j - a_k *and* a_l - a_m - a_n), the solution at $time_0$ shall contain, as subordinate sequences, the solution at subsequent nodes considered as independent decision trees. Generalizing and putting into symbols: for every tree or subtree T and R of a choice dilemma, if $T \subset R$, then $S(T) \subset S(R)$.

This second principle is purely formal. It is purely a statement of the formal principle (true or false, as it may be, when it comes to real-life

opportunity structures) that an embedded dilemma is a dilemma in its own right. Adopting this principle within a theory of practical wisdom amount to embracing the formal idea that the solution to an embedded dilemma must be a component of the solution of the dilemma in which it is embedded.

It is worth noting here that this principle of separable choice, as a solution concept, coincides exactly with the game theorist's notion of *subgame perfect Nash equilibrium* (in the multi-player case) pioneered by Reinhard Selten. This point will help us draw certain parallels later on.

Now, are these two principles compatible? This is an intriguing question in its own right. And answering it amounts to deciding how to conceive of the project of individuating dilemmas—of categorizing decision problems. Put another way around, it amounts to deciding which dilemmas or decision problems we want our account to address as if they were identical. Obviously, if the two principles are not compatible—if, that is to say, they lead to divergent (or more precisely, non-overlapping) solution sets of the same dilemma—we cannot combine them. And so we shall have to decide between them, at least should it turn out that no principle more appealing still presents itself. And we shall have to name the grounds for our preference among all eligible principles, in any case—to justify a choice among competing solution principles.

Now, some writers defend separable choice, others reject it on the grounds that it violates the (purportedly self-evident) principle that we must always aim at the best.[3] There is some reason to think that the principle of separable choice is required to allow for working backwards: if, at the initial point of deliberation, the agent facing a dilemma can foresee what course of action will rationally be chosen at a future node, this might provide a reason for working out what to choose on nodes closer to the point of deliberation.[4] Game theorists refer to deliberations

3 Bratman (1987, 1999) defends it, and Gauthier (1994, 1997) and McClennen (1990, 1992) reject it.

4 McClennen (1992, 1990) gives no attention to this matter.

that incorporate such lines of reasoning, as *backwards induction*. For a trivial example, suppose that I know I will prefer (for logistic reasons) not to drive to the grocery tomorrow, but that I must secure groceries either today or tomorrow, this might be a decisive reason to go today. But this form of reasoning can't go through unless I suppose that the solution to my dilemma of whether to go the grocery today or tomorrow, must overlap with the solution to the (possibly hypothetical) choice I shall face tomorrow once I've settled the issue today. This way, the principle of separable choice makes certain solution or solution procedures possible. But it conflicts with the principle of aiming at the best option. Here is how.

Toxin

A representation of Gregory Kavka's toxin puzzle results if we conceive of the structure depicted in Figure 1 as a dilemma for a single individual who has to take action at both nodes 1 and 2. The agent in question has received a credible offer that pays one million dollars tomorrow morning if, at midnight tonight, the agent manages to produce a genuine intention to drink a vial of vile and sick-making, but ultimately harmless toxin tomorrow afternoon. The agent receives the prize whether or not the toxin is actually drunk tomorrow (Kavka, 1983). The action called for at node 1 is the formation of an intention (in this case, to drink a toxin: a_1 intends and a_2 does not intend), and the action called for at node 2 is the action that is the object of the original proposed intention (drinking the toxin; a_3 drinks but a_4 does not drink). We stipulate furthermore that the outcomes are ranked by the agent as follows (where 'A$\succ$B' means 'A is preferred to B' and 'A$\succ\succ$B' means 'A is much preferred to B'):

(Toxin): $o_1 - o_4 \succ o_1 - o_3 \succ\succ o_2$

(Toxin) says that the outcome of forming an intention to drink the toxin and then proceeding with the target action is much preferred to the outcome of the course of action of never forming the intention in the first place, but that the outcome of forming the intention and then not

following through is most preferred of all the outcomes available. Figure 2 presents the dilemma in node form:

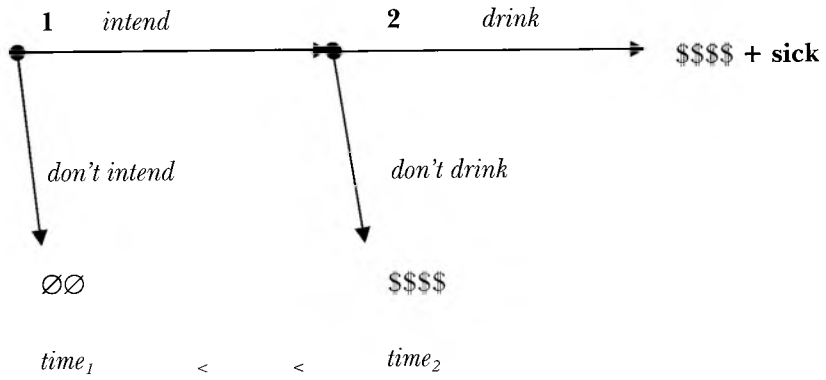

Figure 2: Toxin Puzzle

On Kavka's own view, the compound outcome $o_1 - o_4$ (of action path $a_1 - a_4$: intend but don't drink) is not rationally realizable, with full information of the situation. And his conclusion is based on the principle (we might like to call it *Kavka's principle*) that it is irrational to form an intention (plan) to perform an action that one knows ahead of time one will have no reason whatever to follow through with (put slightly differently: one cannot formulate a plan that one knows in advance one will want to rescind or abrogate). And he held moreover that, this being the case, $o_1 - o_3$ (intend and drink) is likewise unrealizable for the rational agent. David Gauthier accepts Kavka's principle, but denies that $o_1 - o_3$ (intend and drink) is, therefore, unrealizable. He maintains that it is precisely because $o_1 - o_4$ (intend but don't drink) is genuinely unrealizable, that $o_1 - o_3$ (intend and drink) is indeed the most rational of the agent's options. And that this shows something very fundamental to the rationality of deliberation—namely, that the two principles I've enumerated cannot always be satisfied together, and that the rational person must rank the principle of separable choice *below* the principle of maximal outcome.

Degrees of freedom

The arresting—and trying—features of the toxin puzzle are better articulated by a slightly different case, representable again by the structure of Figure 1. In this case (Figure 3) we suppose that a certain agent 1 chooses at node 1, but that a second agent—agent 2—chooses at node 2. And suppose once again that

$$o_1 - o_4 \} o_1 - o_3 \} \} o_2$$

(where we can think of the two agents as agreeing on these preferences; suppose that they will share $2M upon successful completion of a_1, and only one of them—prearranged in advance by a random device—is appointed to drink a poison goblet). Here the causal gap[5] between a_1 (intend) and a_3 (drink) is enormously widened. Without changing the relevant features (features of the structure of preferences and the time sequencing) of the dilemma, we have added degrees of freedom by multiplying sites of agency.[6]

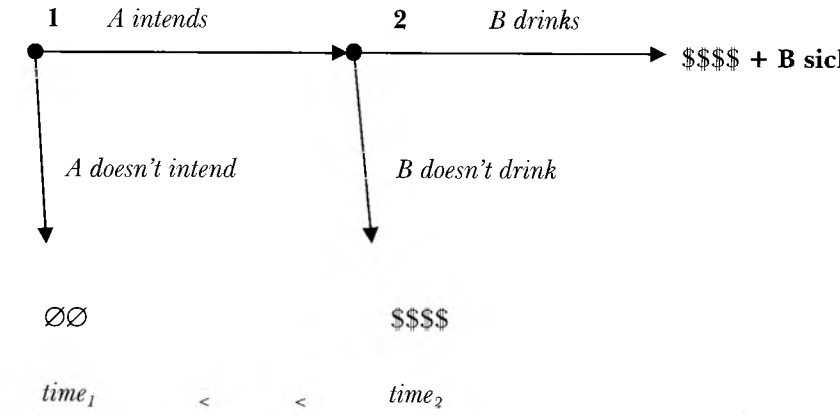

Figure 3: Revised Toxin Puzzle

5 I am on record as not approving of causal language in connection with the events leading up to an action (Thalos 2007) but here it will save on effort to be terminologically sloppy.
6 Precisification of this idea, with applications to human sciences, is in Thalos (1999).

In the modified scenario, the credentials of the claim to the effect that a_1 - a_3 (A intends and B drinks) is unrealizable, are considerably better (a fact due to distribution the degrees of freedom over more sites of agency). And this now gives prominence to the crucial question of whether a_1 - a_4 (A intends and B doesn't drink) is realizable—and if not, why not? This issue turns simply on whether intention formation, as such and simply on its own, can be a target of action in its own right. Can a rational agent form an intention for its own sake, and anticipate its demise without its having done any further work in the economy of said agent's corpus of actions? In other words, can intentions, as such, be the targets of deliberation?

I will leave this (very interesting) issue without attempting an answer. Note, however, that it is a much narrower issue than routinely taken to be. It is a question of what can be the target of an episode of deliberation. Does it have bearing on questions of whether it might be advisable or wise to follow through on a threat or promise? This is questionable (*contra* Gauthier 1994). And much less obvious is the bearing of this matter on the further matter of whether it is advisable or wise to *issue* threats or *make* promises. And much more questionable still is its bearing on whether an agent ought to adopt a *policy* of following through on threats or promises. A great deal more by way of analysis is required to illuminate the bearing of the initial issue (of whether intentions can be targets of deliberation) on further practical, moral and policy matters.

I collect in this paragraph some preliminary lessons of our inquiry thus far. First: the idea that some capacity for planning makes one eligible of certain types of moral consideration is very questionable, if it turns out that for example the Venus flytrap qualifies as having planning capacities on a continuum with those displayed so prominently by primates. And if we should like to retain, at least as a postulate, that the capacity for *complex* planning of the kind of which at least some humans are capable at the zenith of their mental powers, renders an entity eligible of certain kinds of treatment, we shall have to justify this

postulate by offering some reasons for rejecting the contrasting proposal that the capacity for *simple* planning—the capacity to have one's efforts organized over time in a simple way as a matter of Nature's outright gift—does not project the same moral profile. Second: the contention that analysis of planning can bear upon the demands of rationality too is questionable, and for the same reason: many entities are eligible of membership in the category of planners that don't seem eligible of membership in the category of conceptualizers or thinkers at all, let alone rational ones. The multitude of ways and means by which one's efforts can be organized over time, and the fact that different organisms can possess one form of organization (or even several) but not others, suggests that this (very organic) quality of being possessed of organization may be quite distinct from, but continuous with, a capacity to reason about choice.

Finally, the proposals on which our observations cast some doubt, are got by taking an overly intellectualized conception of planning, and furthermore one that (as we learn by and by) makes it impossible to study planning, as a feature of agency, as a natural phenomenon in the biological regime. Accepting such an intellectualization of planning makes it seem as if being organized in one's efforts is not part of the natural order but in some way transcends it. And this is obviously incorrect. And when practical wisdom is conceived as something that transcends the natural order of things, we lose the ability to integrate our considerations—moral and otherwise—for human agents, with our considerations for nonhuman entities.

Larger morals of the exercise

First important moral: calling upon principles alone, such as those displayed above, and turning the searchlight inward upon intuition for their affirmation or disaffirmation, does not result in resolution of any substantive issues regarding what is called for in practical wisdom. One reason is that this procedure cannot by itself settle important matters with bearing on the substantive questions—such as the number of

degrees of freedom in any given true-life scenario conforming to the bare-bones event structure in any given figure. For neither counting sites of agency, nor counting nodes or forks, gives an accurate count of the degrees of freedom. This is illustrated by the fact that, as we saw, Figures 2 and 3 have the same opportunity structure, but not the same agentic structure. Agentic structure cannot be captured by examining formal features only of a decision problem. And this fact helps illuminate why refinement of the Nash equilibrium solution concept, that we mentioned earlier in connection with separability, has now been abandoned by a preponderance of game theorists.[7]

Further compounding this problem is the fact that the number of degrees of freedom is (at least sometimes) under the control of the entities occupying the sites designated or suggested in the figures: it is sometimes an agent's prerogative to relinquish control via acquiescing in a certain control protocol, as we will discuss shortly. This is not to say that the agents in question do this voluntarily, or even that they relinquish responsibility for the outcome, as we will now discuss.

It might help here to make some further remarks about agentic structure, and the correlative notion of a degree of freedom. We have been working with the idea that (at least) one defining characteristic of agency is aggregatability: so entities on one scale of organization can come together under some further 'rules of engagement' or 'protocol' to form a (single) agent at a still higher scale, and likewise a higher-scale agency can disintegrate into numerous smaller-scale agencies. Agencies both coalesce and dissolve.[8] Agency is thus, on the conception we are working with, a matter of integration, and this is in turn a matter of organization. A degree of freedom, in the sense we will use the term here, does not amount to freedom in the ordinary sense—it is not simply

7 Don Ross, personal correspondence, puts the point this way: 'there is no perspective in most games that embodies and consolidates into coherence the full cascade of interlocking agents' beliefs'.

8 Thus far I have defended this thesis only piecemeal in articles scattered across a range of journals, but a more coherent defense of this thesis is in progress.

the capacity for putting some plan into action. Freedom in our (technical) sense, and as physicists and engineers too use the term, is the *absence of (some) constraint upon magnitude*.[9] And the *number* of degrees of freedom is thus a count of constraints that must be placed on a system to achieve a full determination (or as full as nature will allow) of the magnitude of all its unconstrained quantities—in our case, behaviors. This is a Systems conception of freedom, because the number of unconstrained quantities is system-relative. This conception rests on the following ideas.

> ***Definition DoF.*** **X** is an *independent quantity* or a *degree of freedom (dof) of a system* σ, *according to a theory or scheme of representation* $T =_{df}$ **X** is named or otherwise designated by T as belonging among those quantities whose magnitudes shape the state of σ.

In this definition the notion of shaping is used primitively, and is taken as governed by the following axiom:

> ***Axiom DoF.*** If a quantity **X** shapes the state of a system σ, or if it gives shape to a quantity **Y** of σ, then it is false that **X** is given shape to by any other quantity.

We can apply these ideas in decision situations, as follows:

> ***Definition Decision DoF.*** Some factor or entity (for example, a choosing 'atom' or 'molecule') is a degree of freedom in a deciding/acting system S, according to decision theory DT, if and only if it is among those factors designated by DT as shaping the behavior of S.

Now, when we count the number of degrees of freedom in, for example, a decisional system—a system in which the outcome depends on a (potentially interdependent) network of decisions—we are positively *not*

9 For a great deal more on this contrast, as well as how we shall have to refine the conception when we bring together all the sciences into a coherent picture, see Thalos (1999) and 'Nonreductive Materialism Without Hierarchy,' in progress.

looking at the capacities for action on the parts of decision makers involved. We are looking, instead, at the numbers of uncontrolled, unconstrained or unknown decision-relevant factors that can swing the system towards one set of behaviors rather than another. Contrariwise, the degree of *control* in such a system is directly proportional to the exercise of what we may refer to as *social functioning* within that system—the exercise of (learned) capacities for smooth social and decisional interaction (social 'flow').

Recall the example of flocking we discussed above. That example too can be handled very well with these conceptual tools. Flocking needs to be very finely-tuned to achieve a particular end. Close behavioral coupling among near neighbors in a flock allows a localized change in direction to be amplified and propagated across the flock. This allows each flock member to influence and be influenced by flockmates much farther away than their local neighborhood—it gives each a much larger 'effective perceptual range' than their actual sensory range. This scaling is nonlinear. Study of the details of the scaling relations reveals that it is hard for groups to maintain cohesion if the coupling distance is too short. Longer-range transfer of information is enabled by increasing the coupling distance. Increasing the coupling distance further still creates a cohesive group, but 'misinformation' might be propagated (as use of information about motion of distant individuals is in some circumstances less beneficial locally).

In addition, coupling may be moderated by context-dependence. For example, if individuals conditioned reactions upon context (under threat, for example, aligning more strongly with distant flockmates, thereby increasing 'system gain'), this could allow for some flexibility, but there is a cost. Heightening sensitivity to weak or ambiguous environmental signals increases susceptibility to 'false positives', just as damping response to local fluctuations in less threatening contexts increases 'false negatives'. And so a balance has to be struck:

> Under different circumstances individuals may adopt behaviour that facilitates collective damping of local fluctuations. During long-distance migration, for example, animals are often faced with the challenge of navigating up noisy and weak thermal or resource gradients. Local variability makes this task difficult, or even impossible, for individuals in isolation. But coherent interactions can allow groups of organisms to function like an integrated self-organizing array of sensors, again increasing effective perceptual range. As long as intereactions are sufficiently sensitive to ensure cohesion, but not too sensitive to local fluctuations and individual error, individuals can effectively respond to the weak long-range gradient (from Couzin 2007; *cf.* Couzin and Krause 2003, as well as Couzin, *et. al.*, 2005).

And so a flock of hundreds of organisms, operating under a given set of 'rules of engagement', is decidedly *not* a system with degrees of freedom on the order of hundreds or more (as a count of the behaving 'sites' would have suggested): it is instead a system with something on the order of a dozen degrees, counting among them rough size, coupling distance, and level of context-sensivity, as well as environmental variables that tend to couple with these features. A flock is an entity with a reduced number of degrees of freedom than there would be without the rules of engagement. And these degrees of freedom will be revealed as we study (and model) the dynamics of their behavior. It is not a matter of decisional nodes, forks or sites.

Bringing this point back home: humans too (specifically in their capacities as agents) operate with 'rules of engagement', some more local than others. And these tend to reduce the number of degrees of freedom from what they would have been without the 'rules'. Reductions in degrees of freedom between rational decision makers, for example in the interests of coordination, moves the decision 'atoms' closer in the direction of a flock. And which variables function as *the* degrees of freedom will vary according to the specifics of the system (whether, among many things that matter, it is composed entirely of adults, or equals, and how far from each other they might be).

But what *can* settle the substantive matter as to the number of degrees of freedom? There are three ways to go with this question:

1. ***Assume*** *that to every body/center of consciousness there corresponds one degree of freedom or agent;*

2. ***Decree*** *that the situation (the opportunity or event structure) shall settle the question of the number of degrees of freedom, given as collateral information or 'boundary conditions';*

3. ***Assume nothing****, allowing that even the number of degrees of freedom could be a matter to be determined by the solution process.*

How to proceed with this question is profoundly important to the analysis of agency. The first way simply announces *ex cathedra* an individualistic doctrine on agency. (It is typically inspired by a transcendental stance on the topic.) The second way takes the facts of agency to be predetermined in advance by possibly empirical, but ultimately non-strategic facts on the ground. The third way allows agency to be negotiable within the 'decision game', the 'game of practical wisdom.' (Decision theorists originally thought that these too are games, and could be appropriate handled endogenously, but they are coming to realize that not all matters can be settled endogenously. Hence the abandonment of Nash Equilibrium program.)

And so the choice among these three can hardly be settled *a priori*. Still, the choice is important to the metaphysics of agency. Analysis of practical wisdom makes progress only in proportion to how well it conceptualizes the metaphysics of agency, and taxonomizes such agencies as it encounters in nature in ways that go substantially beyond what is given in the event structures displayed in these abstract figures, in true-life problems faced by true-life, flesh-and-blood agents.

The second important moral: selection of methodology (3) above goes hand in hand with making room for an evolutionary account of the units of agency. In other words, it leaves room for exploring whether evolutionary forces can have had some impact upon the constellation of agency structures we find on the ground, whereas selection of methodology (2)—and obviously of (1)—waives all attempts at such an exploration.

The third and perhaps most important moral: the units of agency issue—the question of whether there is a unit of agency that is most

fundamental or independent, such that other units of agency are in some way dependent upon it—is tightly interwoven with questions concerning the nature and types of goal attainment/planning strategies, and how they might emerge in the course of natural history. And in that way both questions bear on the choice of principles for rational decision we have briefly reviewed above. For example, if it should turn out that front-loading is the most vigorous, most effective or most efficient of all goal attainment strategies, we should expect to find more 'local' than 'distributed' agency structures in nature. And accordingly we should also accept a certain principle of separable choice as a principle of rational decision.

I favor methodology 3. Not only does it make room for exploration of questions about the units of agency and goal attainment strategies, within an evolutionary setting, but it also urges these questions as important and fundamental to an inherently interdisciplinary enterprise of studying agency. It prompts for a natural history of the units of agency we find in nature. And it makes it clear that a natural history is both necessary and desirable for purposes of an account of moral agency.

Principles, again

What support can we offer the principles of choice (maximization or separability) displayed above, short of appeals to intuition or to other principles that might support them?

One untried approach is to inquire what practical wisdom might dictate in those instances where foresight (full information) is unavailable—where for example long-term consequences cannot be clearly predicted. So, for example, rather than focusing upon the T_2 of $T_{1\text{-}2}$, what if we should focus instead upon the T_1 of $T_{1\text{-}2}$. Is there a working-forward principle, as contrasted with a working-backwards principle? It would be extremely useful to work out such an account, for a host of reasons:

1. There is very often precious little to go on in the way of long-term information.

2. Mistakes are inevitable. How does one move forward from having made a mistake within a plan?

3. Not only are mistakes inevitable, but so is maturation. We are entities that follow well-established developmental trajectories. How do we accommodate the inevitable yet uncertain changes that will come, particularly when at the original stage of development there is precious little in the way of practical wisdom to support them?

Another approach is to take the structures one at a time, and to determine whether there might be principles that govern structures much more individually. This is a local, as contrasted with a global, approach. So, in the structure of Figure 1, we might propose this, as a principle of high generality:

If $o_2 \nmid\nmid o_1 - o_3 \nmid o_1 - o_4$, then $S(T_{1\text{-}2})=\{a_2\}$

no matter how many sites of agency are in play. This principle might seem to follow from the principle of maximal outcomes. (But perhaps it does not follow straightforwardly, since in this principle we are aggregating outcomes $o_1 - o_3$ and $o_1 - o_4$.) And we would have to think the matter over carefully, and for different numbers (and kinds) of sites of agency, once we departed from this very simple structure.

Metaphysics and development: the tangled webs we weave

The reasons enumerated in the last section, in favor of seeking forwards-working principles, indicate that a good deal has been left out of the analysis of practical wisdom and planning thus far, for example:

(1) Who is the agent, and how do the dilemmas hang together for them (how far ahead can they project their decision trees)? This matter concerns issues surrounding how to represent the information at their disposal faithfully;

(2) What developmental stage have the agent 'sites' in question achieved? This question has bearing on the matter of the level of organization within their reach; and this, in turn, will have a bearing on what combinations of actions are realizable. If for some (highly mature) agents, certain combinations of actions are unrealizable, it is similarly true that for other (much less mature) agents, other—different—combinations are unrealizable. For example, some agents do not respond to threats, however credible, just as they do not respond to promises either; this is not *obviously* a failure of any kind, and a failure of rationality in particular, especially when we take into account the fact that public knowledge of such incapacities can, and often does, work in favor of such an agent.[10] And too whether it will work in their favor depends on who they are interacting with, as much as what is at stake for them and for others. Similarly, some agents (notably, the younger set) do not follow easily through with promises, threats or even plans. And this can certainly work in their favor (as every parent knows). The toxin puzzle is a case in point: an agent so organized as to be capable of forming genuine intentions that they can foresee rescinding, will be in the happy position of being able to realize $a_1 - a_4$

By the same token, one can ask whether there are considerations besides outcomes that are taken into account by the agent or agents in question, and how these considerations, when taken into account by the relevant agents, bear on the questions we have already canvassed. So, suppose that someone is moved by (so-called) moral considerations in favor of promise-keeping. Does this create a situation in which certain combinations of actions are in a relevant sense not realizable for them?

10 Schelling (1960) makes a good deal—and good use—of this point.

And by the same token, suppose that certain other folks are not moved very much by such considerations. Might this not open up options for them that will not be open to other agents under the very same circumstances?

It is not at all obvious what we must say about such matters, from the stance of a theory of goal attainment and deliberation. But these questions are, and now clearly, pertinent.

Toxin, again

Thomas Schelling famously brought such questions as we have just reviewed to prominent attention. Others—like Edward McClennen and David Gauthier—have sought to address such questions. All without articulating any account of the bearing of metaphysics of agency on decision or other organizational questions. They have sought to take metaphysics off stage. And their accounts suffer for it: easy counterexamples to their principles can be had simply through changing and juggling degrees of freedom.

One way of seeing that this stratagem of altering the degrees of freedom will bring any principle down, is to notice how it can easily take down even the principle of maximal outcomes. Again take the structure of Figure 1 and suppose that:

$$o_1 - o_4 \} o_1 - o_3 \} \} o_2.$$

The principle of maximal outcome would insist that $S(T_{1\text{-}2})=\{a_1 - a_4\}$. And we know that a very worthy counterexample is Kavka's very own toxin case: a case with that structure but where the degrees of freedom between $a_1 - a_4$ is reduced to something less than the 2 that the structure would suggest.

Someone might be tempted to complain that this counterexample works simply because it shows that a certain stratagem is not realizable. And so it works by displaying that the structure of Figure 1 is an inappropriate representation of the decision problem. But this complaint only serves to provoke the response that my criticism can be put in different terms, without altering its bite. This is that decision trees, as such, are—contrary to how they are often depicted—inadequate representations of the true-life dilemmas: trees leave out the all-important metaphysics. But this is the very same criticism as mine, put in different words.

Metaphysics, the final frontier: The dimensions of agencies

How do sites of agency coalesce? They do so along a variety of different dimensions. The first and most obvious is that agency can bring together control over a certain array of resources; when this occurs so, we can speak of a *pool* of resources. I will refer to this as the *zeroeth* dimension of combination or merger. And I will offer (what I like to think of as) a Lego model of agent formation. The zeroeth dimension is the emergence of a unit Lego block.

Second, and also very familiar, sites of agency can combine along different life slices and developmental stages of the same organism; when they do, we speak of an *individual*. This I will refer to as the vertical dimension of merger, conceiving of time as advancing upwards. Individuals will be represented by towers consisting of unit pieces.

Finally, and most controversially, sites of agency located in different organisms can coalesce (among members of the same species, but also across species boundaries, as when a human being trains a companion animal for hunting or herding); when they do, we speak of *coalitions*. I will refer to this as the *horizontal* dimension, conceiving of coalitions as overlaps or unions in the horizontal plane. In the Lego model, a coalition is a multi-unit block that joins more than one tower. Enduring coalitions will consist of multi-unit towers rising vertically from the point of merger.

The natural history of alliance and merger will seek answers to questions of the form: Are alliances more fit than single towers? How do short-term alliances fare against longer-lived ones? Which are the most stable forms and patterns of merger? How do networks of alliances interact with other networks? Are there universal properties we should

expect to see in large networks of alliance? And what happens when we move up the size scale? Do we at some point get a 'phase change'—do we encounter points at which 'phase' changes, conceived as quantum changes in some behavioral variable, occur? These questions, obviously, are beyond the scope of this essay, but they suggest a new realm of scientific inquiry that has been sorely neglected.

Acknowledgements

It is a pleasure to acknowledge the generosity of kind colleagues in my home department and in the academy at large: Chrisoula Andreou, Leslie Francis, Lije Millgram, Don Ross and an anonymous referee for this journal—who read and commented on early versions, lending their good judgment to the tempering of mine. None of them, however, is to be held to blame for any excesses still clinging to my execution of this project. But special thanks to Don: he admonished me for failure originally to draw parallels between (on the one hand) my points about the nonidentity of tree structure and true-life problem structure, and (on the other hand) game theorists' recent recognition that the Nash refinement program has to be abandoned.

University of Utah

References

Bratman, Michael, 1987. *Intention, Plans, and Practical Reason*. Cambridge: Harvard University Press.

Bratman, Michael, 1999. *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge: Cambridge University Press.

Bratman, Michael, 2000a. Reflection, Planning, and Temporally Extended Agency, *Philosophical Review*, 35-61.

Bratman, Michael, 2000b. Valuing and the Will, *Philosophical Perspectives*, 249-265.

Bratman, Michael, 2002. Hierarchy, Circularity, and Double Reduction, in S. Buss and L. Overton, eds., *Contours of Agency: Essays on the Philosophy of Harry Frankfurt*. Cambridge, Mass.: MIT Press, 65-85.

Couzin, Iain, 2007. Collective minds, *Nature* 445, 15 Feb 2007, 4.

Couzin, I. D. and J. Krause, *J. Adv. Study Behav.* 32 (2003), 1-75.

Couzin, *et al.*, 2005. *Nature* 433, 513-16.

Gauthier, David. 1997. Resolute Choice and Rational Deliberation: A Critique and a Defense, *Nous*, 31(1), 1-25.

Gauthier, David. 1994. Assure and Threaten, *Ethics*, 104 (4), 690-721.

Hardcastle, V. G. 1995. A Critique of Information Processing Theories of Consciousness, *Minds and Machines*, 5, 89-107.

Kavka, Gregory. 1983. The Toxin Puzzle, *Analysis*, 43, 33-36.

McClennen, Edward F. 1997. Pragmatic Rationality and Rules, *Philosophy and Public Affairs*, 26 (3), 210-258.

McClennen, Edward F. 1992. Rationality and Dynamic Choice, *Ethics*, 103 (3).

McClennen, Edward F. 1990. *Rationality and Dynamic Choice*. New York: Cambridge University Press.

Norman, D. A. and T. Shallice, 1986. Attention to Action: Willed and Automatic Control of Behavior, in *Consciousness and Self-Regulation: Advances in Research and Theory*, edited by Richard Davidson, Gary Schwartz and David Shapiro. New York: Plenum Press.

Schelling, Thomas. 1960. *Strategy of Conflict*. Cambridge: Harvard University Press.

Thalos, M. 2007. Sources of Behavior: Towards a Control Account of Agency, *Distributed Cognition and the Will*, Don Ross *et al.*, eds., Harvard: MIT Press, 123-67.

Thalos, M., 2002. Explanation is a Genus: On the Varieties of Scientific Explanation, *Synthese*, 130, 317-354.

Thalos, M. 1999. Degrees of Freedom in the Social World. *Journal of Political Philosophy*, 7, 453-77.

Wegner, Daniel, 2002. *The Illusion of Conscious Will*. Cambridge: MIT Press.